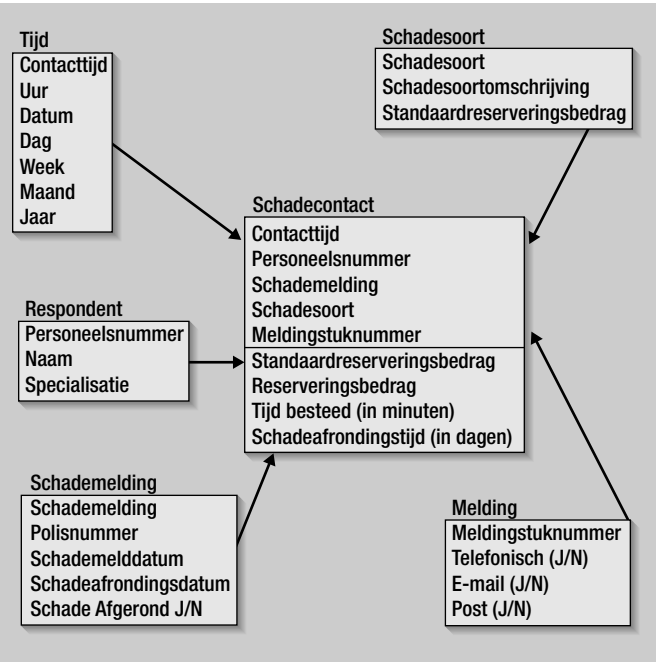
Problemen oplossen binnen een 'conformed dimensions' model

Een fact is een fact als er fact op staat

Karien Verhagen

U kent ze wel, de modelleringscursussen voor en discussies over de specifieke datamodellering van Datawarehouses. Tien manieren om de historie op te lossen, al of niet platslaan, de junk- en minidimensions en alle varianten en combinaties. Gewapend met de kennis uit de cursus of uit boeken gaan vervolgens de datawarehouse-architecten aan de slag. Wat blijkt? Voor het eerste praktijkprobleem voldoen de tien aangeboden oplossingen niet en een elfde wordt gecreëerd.

Wanneer vervolgens ook nog eens rekening gehouden moet worden met niet *één* maar *meer* bestuurlijke informatiebehoeften in *één* model, blijkt het toch steeds weer een illusie dat het mogelijk is om alle huidige en toekomstige problemen in *één* laag op te lossen. Kubussen of sterschema's koppelen door *conformed dimensions* blijkt lang niet altijd mogelijk. Kubussen aan elkaar knopen: wel eens geprobeerd?



Afbeelding 1: Het vraaggerichte model.

INTEGRATIEPROBLEMEN

Het komt steeds minder vaak voor dat bedrijven van scratch beginnen met een datawarehouse. De uitdagingen waar de datawarehouse-specialist mee te maken krijgt in de praktijk, zijn niet meer de problemen van de sponsoring en de time-to-market van een eerste increment. Het zijn steeds vaker de integratieproblemen. Een belangrijk integratieprobleem is het koppelen van losse datamarts, al of niet door flitsende boys van buiten geïmplementeerd.

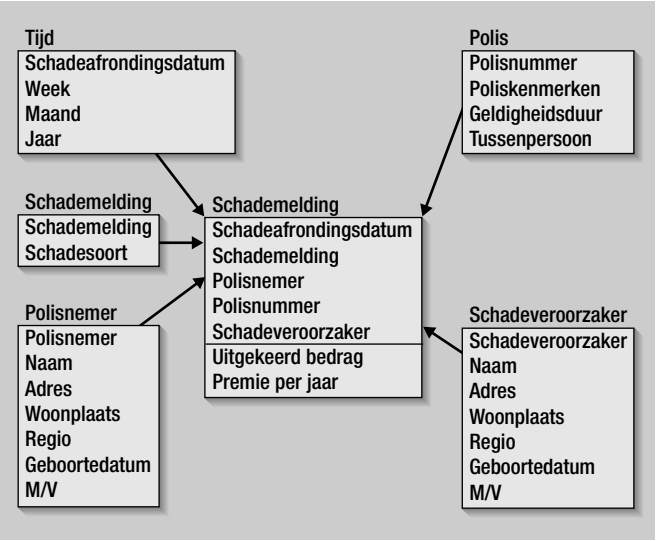
Wanneer je constateert dat drie datamarts van dezelfde gegevens gebruik maken, wil je uiteraard de consistentie bewaken en het liefst die gegevens uit eenzelfde bron betrekken. Er kan een informatievraag komen die gebruik maakt van gegevens uit een dwarsdoorsnede van twee van je datamarts. "Conformed dimensions" zegt Kimball dan, maar dat blijkt lang niet altijd te kunnen. Dat kan bijvoorbeeld niet als het fact ineens een dimensie blijkt te worden in een andere datamart.

Dit artikel illustreert dit probleem met een praktijkvoorbeeld uit de verzekeringswereld. De verzekeringsmaatschappij in kwestie heeft in de loop van de tijd drie informatievragen met een sterschema-oplossing beantwoord.

EVALUATIE RESPONDENTEN

De schade-afhandelingsafdeling van een verzekeringsmaatschappij behandelt claims op schadeverzekeringen. Het betreft verschillende soorten schades. De respondenten handelen deze schades af. Elke schadeverzekering heeft een aantal schadesoorten (voor autoschades bijvoorbeeld met of zonder letsel). Voor elke schadesoortcode bestaat een standaard reserveringsbedrag. Dat is voor de boekhouding het bedrag dat wordt uitgekeerd en gereserveerd tot het tegendeel bewezen is. Aan de hand van nieuwe informatie (per brief, telefoon of e-mail) kan het reserveringsbedrag worden bijgesteld. Met de afhandeling van schades is tijd en geld gemoeid.

De eerste informatievraag luidt: de schade-afhandelingsafdeling wil de kosten en de bestede tijd evalueren. Ook wil ze de correctheid van het standaard reserveringsbedrag graag evalueren, niet



Afbeelding 2.

alleen aan de hand van het uiteindelijk uitgekeerde bedrag, maar ook aan de hand van de tussentijdse bijstellingen. Tussentijdse bijstellingen kunnen plaatsvinden omdat nieuwe informatie over het schadegeval bekend wordt.
Per schademelding wil men weten:

- Hoe vaak is er contact geweest voordat de schade is afgerond en hoeveel tijd is in totaal aan de schademelding besteed;
- Hoe vaak is het reserveringsbedrag aangepast;
- In welke mate wijkt het reserveringsbedrag af van de uiteindelijke uitkering en de tussenschattingen;
- Is er vaak telefonisch contact geweest of ging de afhandeling via post of e-mail;
- Is hier een verschuiving in te constateren;
- Hoe staat de tijdsbesteding van de respondenten in verhouding tot de kosten van de schademeldafdeling;
- Zijn afwijkingen causaal te traceren naar het soort schade, de respondent, het contactmedium, de schadesoort.

Het (vraaggerichte) model dat in deze informatiebehoefte zou kunnen voorzien, is weergegeven in afbeelding 1. (De termen Vraag- en Aanbodsgericht model zijn voor zover ik weet voor het eerst gebruikt door Harm van der Lek en door mij dankbaar overgenomen.)

Over de modellering valt te twisten. Het standaard reserveringsbedrag is opgenomen als 'platgeslagen' historie in de feitentabel. Dat wil zeggen: als het reserveringsbedrag in de tijd zou veranderen, verandert het alleen in de dimensie, niet in het feit. De schade-afrondingstijd is een procesgegeven en geeft de totale doorlooptijd tussen melding en afrondingsdatum. Deze modellering is gevaarlijk omdat de granulariteit per schadecontact en niet per schademelding is bepaald. Het wordt gevuld bij de finale afronding, bij alle tussentijdse meldingen staat het veld op nul.

De tweede informatievraag: de schade-analyse-afdeling van diezelfde maatschappij wil de uitgekeerde schadebedragen analyseren ten opzichte van de geïnde premie en afwijkingen kunnen traceren naar de demo- en geografische gegevens van de polisnemer, de veroorzaker van de schade (als bekend), de schadesoort en de tussenpersoon en ongetwijfeld nog veel meer kenmerken die gemakshalve voor deze case weggelaten zijn.

Het model dat in deze informatiebehoefte zou kunnen voorzien, staat in afbeelding 2. Ook hier weer de opmerking dat premie per jaar maar een keer gevuld mag zijn omdat de granulariteit op schademelding is. Er zijn polissen met meer schademeldingen per jaar. Er zijn ook polissen zonder schademeldingen. Er kan bijvoorbeeld een null-meldingsrecord worden gemaakt waarin de premie per jaar wordt gevuld, zodat de measures additief worden.

De derde en laatste informatievraag betreft de evaluatie van de tussenpersonen:

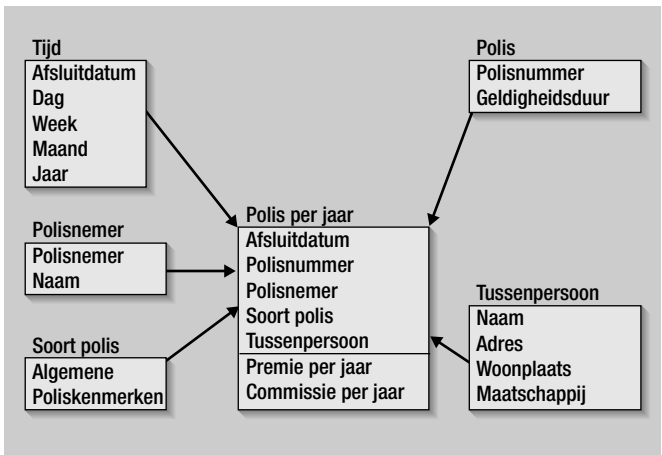
- Hoeveel commissie ontvangen ze in verhouding tot de geïnde premie? Hoe is het verloop daarin;
- Hoeveel nieuwe polissen zijn er afgesloten en is er een causaal verband naar de soort polis;
- Welke tussenpersonen zijn actiever dan anderen.

Een model dat zou kunnen voorzien in deze informatiebehoefte is te zien in afbeelding 3.

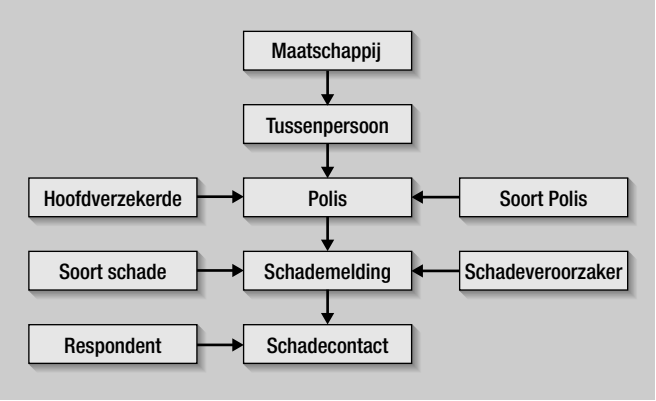
HOEZO CONFORMED DIMENSIONS?

Als deze drie informatiebehoeften moeten passen in een 'conformed dimension' model ontstaat een probleem. De facts die in het ene model staan afgebeeld worden dimensies in het andere.

Uit het rijke scala aan oplossingen dat de 'Lifecycle Toolkit' van Ralph Kimball biedt, zie ik geen oplossing waar dimensies in facts veranderen en omgekeerd. Toch zouden alle drie de informatiebehoeften in te vullen zijn in datamarts vanuit een centraal model, dat er bijvoorbeeld uitziet als in afbeelding 4. Of dat relationeel of OO moet zijn, in hoeverre je hier al platslaat, optimaliseert en denormaliseert ? Je kunt er lang over debatteren, maar hoe interessant ook, dat was hier niet de vraag.



Afbeelding 3: Polissen kunnen in 'pakketcombinaties' en met speciale voorwaarden worden afgesloten. Deze kenmerken staan in 'Soort polis'.



AFBEELDING 4: HET CENTRALE MODEL.

CONCLUSIE

Wanneer het centraal model een performante view of een virtuele datamart kan leveren die in de informatievragen 1, 2 en 3 kunnen voorzien, heb ik de daadwerkelijk opslag beperkt tot één centrale opslag. Bovendien ben ik gewapend tegen vragen naar dwarsdoorsneden. Denkbaar is bijvoorbeeld een relatie tussen tussenpersoon en frequentie van schadecontacten. Kortom: ik pleit voor de 'Inmon-oplossing' met een centraal datawarehouse en indien nodig om redenen van semantiek of performance, een uit het centraal model onttrokken relationeel model of een OLAP-kubus. Dat kan een view zijn of een hard extract met bijvoorbeeld aggregaten, de term Datamart is daarvoor gebruikelijk.

Het centraal datawarehouse is een aanbodsgericht model. Dit model is een afspiegeling van de bedrijfsprocessen en wordt met beleid en incrementeel ingevuld. De vraag aan Ralph Kimball is natuurlijk hoe dit probleem op te lossen zou zijn in een 'conformed dimensions' model. Jammer dat hij DB/M niet kan lezen! Kom, ik zal het artikel vertalen, het hem opsturen en u de reactie laten weten. ●

Drs. C. Verhagen is programmamanager Datawarehousing bij de Bloemenveiling Aalsmeer (VBA).

ORACLE 9I KLAAR VOOR 32- EN 64-BIT WINDOWS SERVER 2003

Oracle heeft haar 9i database release 2 direct aangepast voor 32- en 64-bit Windows Server 2003. Gebruikers die hun Oracle databaseservers willen upgraden naar de nieuwste versie van Microsoft Windows krijgen nu de beschikking over een krachtiger database met meer geheugen en een efficiëntere en gebruiksvriendelijker ontwikkelomgeving. Met een lage TCO vormt deze database voor kleine en middelgrote bedrijven volgens Oracle een aantrekkelijk alternatief voor de Microsoft SQL server op Windows.
Voor meer informatie: www.oracle.com

INFORMATICA INTRODUCEERT BUSINESS ANALYTICS VOOR LINUX

De gehele Business Analytics productportfolio van Informatica ondersteunt nu ook Linux. Hiermee is Informatica naar eigen zeggen de eerste leverancier die een complete end-to-end Business Analytics-oplossing voor Linux aanbiedt. De Linux versies PowerCenter, Warehouse en PowerAnalyzer hebben dezelfde ontwikkelfunctionaliteiten als de versies voor Windows NT en Unix.
Meer informatie: www.informatica.com/nl

BO LANCEERT PRODUCTSUITE ENTERPRISE 6

Business Objects heeft de release aangekondigd van Enterprise 6, een nieuwe versie van BO's BI-suite. Enterprise 6 is een suite van geïntegreerde Business Intelligence producten en bestaat uit bedrijfsanalytische applicaties, een BI-platform en data-integratieproducten. Enterprise 6 bestaat uit nieuwe en vernieuwde producten en vele verbeterde functionaliteiten; naast BO's Warehouse en DataIntegrator omvat de nieuwe release een geheel nieuwe versie van WebIntelligence en een verbeterd BI-portal.
Meer informatie: www.businessobjects.com

CA KOMT MET INGRES VOOR LINUX AMD OPTERON-PLATFORM

Computer Associates International heeft met Advantage Ingres een RDBMS uitgebracht dat gebruikmaakt van de 64-bits-verwerkingskracht van de AMD Opteron-processor. Advantage Ingres moet een naadloze transitie naar de 64-bits Linux-omgeving mogelijk maken.
Het gebruik van 64-bits rekenkracht biedt een hoger prestatieniveau en een grotere schaalbaarheid voor zakelijke VLDB-toepassingen. CA heeft nauw samengewerkt met AMD om een optimale interoperabiliteit te garanderen tussen de CA-oplossingen en de AMD Opteron-processoren. Advantage Ingres-servers met AMD Opteron-processoren kunnen direct een enorm geheugen adresseren. Door het gebruik van grote caches wordt kostbare disk-I/O sterk verminderd en worden tegelijkertijd de prestaties en schaalbaarheid op een hoger niveau gebracht.
Voor meer informatie: http://ca.com

TED CODD OVERLEDEN

Op 18 april 2003 is in Florida Edgar F. Codd overleden aan een hartaanval. Hij was 79 jaar. De van oorsprong Engelse Ted Codd is de grondlegger van het relationele database model, dat hij ontwikkelde als medewerker van IBM, als antwoord op de toen bestaande data management systemen.
In 1981 bracht IBM haar eerste door Codd ontworpen relationele systeem op de markt, SQL/DS; in 1983 volgde DB2. In 1983 raakte Codd ernstig gewond en na zijn herstel verliet hij IBM. Hij bleef doorwerken tot 1999 met de voormalige IBM-medewerkers Chris Date en Sharon Weinberg, met welke laatste hij in het huwelijk trad.